IJSPT

ORIGINAL RESEARCH

# THE INTRA- AND INTER-RATER RELIABILITY OF THE SOCCER INJURY MOVEMENT SCREEN (SIMS)

**Robert McCunn**[1]
**Karen aus der Fünten**[1]
**Andrew Govus**[2]
**Ross Julian**[1]
**Jan Schimpchen**[1]
**Tim Meyer**[1]

## ABSTRACT

***Background/purpose:*** The growing volume of movement screening research reveals a belief among practitioners and researchers alike that movement quality may have an association with injury risk. However, existing movement screening tools have not considered the sport-specific movement and injury patterns relevant to soccer. The present study introduces the Soccer Injury Movement Screen (SIMS), which has been designed specifically for use within soccer. Furthermore, the purpose of the present study was to assess the intra- and inter-rater reliability of the SIMS and determine its suitability for use in further research.

***Methods:*** The study utilized a test-retest design to discern reliablility. Twenty-five (11 males, 14 females) healthy, recreationally active university students (age 25.5 ± 4.0 years, height 171 ± 9 cm, weight 64.7 ± 12.6 kg) agreed to participate. The SIMS contains five sub-tests: the anterior reach, single-leg deadlift, in-line lunge, single-leg hop for distance and tuck jump. Each movement was scored out of 10 points and summed to produce a composite score out of 50. The anterior reach and single-leg hop for distance were scored in real-time while the remaining tests were filmed and scored retrospectively. Three raters conducted the SIMS with each participant on three occasions separated by an average of three and a half days (minimum one day, maximum seven days). Rater 1 re-scored the filmed movements for all participants on all occasions six months later to establish the 'pure' intra-rater (intra-occasion) reliability for those movements.

***Results:*** Intraclass correlation coefficient (ICC) values for intra- and inter-rater composite score reliability ranged from 0.66-0.72 and 0.79-0.86 respectively. Weighted kappa values representing the intra- and inter-rater reliability of the individual sub-tests ranged from 0.35-0.91 indicating fair to almost perfect agreement.

***Conclusions:*** Establishing the reliability of the SIMS is a prerequisite for further research seeking to investigate the relationship between test score and subsequent injury. The present results indicate acceptable reliability for this purpose; however, room for further development of the intra-rater reliability exists for some of the individual sub-tests.

***Keywords:*** Assessment, association football, kinematic, screening

***Level of evidence:*** 2b

**CORRESPONDING AUTHOR**
Robert McCunn
Institute of Sport and Preventive Medicine
Saarland University
Campus Bldg. B8.2
66123 Saarbrücken, Germany
Phone: + 49 681 302 70410
Fax: + 49 (0) 681 302 4296
E-mail: bob.mccunn@me.com

[1] Saarland University, Institute of Sport and Preventive Medicine, Saarbrücken, Germany
[2] University of Bedfordshire, Institute of Sport Science and Physical Activity, Bedford, United Kingdom

## INTRODUCTION

The proliferation of movement screening research and its widespread use in professional soccer reveals a belief among practitioners and researchers alike that movement quality may have an association with injury risk.[1, 2] Movement quality is ill defined but relates to the ability of an individual to perform a given movement in a controlled manner while demonstrating good or acceptable technique. Exactly what constitutes good technique remains a topic of debate. While it is arguable that no 'correct' movement pattern exists for any given exercise there are certain characteristics that may be undesirable, such as restricted range of motion and an inability to control coordinated movements. The rationale behind movement screening is that such limitations may result in acute injuries or contribute to insidious overuse complaints.[3-5]

Numerous screens exist; however, the supporting evidence with regard to both their reliability and association with injury varies widely in both volume and methodological quality.[1] The majority of such research has focused on the Functional Movement Screen (FMS™), which has demonstrated good reliability but conflicting relationships with injury likelihood.[1, 6] The FMS™ was designed as a 'general' movement assessment tool and has been used within a wide range of sports and professional domains including the military and emergency services.[7-9] In contrast, some screens such as the Landing Error Scoring System (LESS) have been designed with the intention of identifying those at an increased risk of a particular type of injury, for example, anterior cruciate ligament rupture.[10] In addition, some have been designed for use within particular sports, for example, netball and rugby union.[3, 11] Despite the popularity of movement screening within professional soccer, no soccer-specific tool currently exists.[2] The present study introduces the Soccer Injury Movement Screen (SIMS), which has been designed specifically for use with soccer athletes. The movements contained within the assessment were selected to reflect the most common sites (lower extremities) and types (strains and sprains) of soccer-related injury and hence they primarily tax the mobility and stability of the ankle, knee and hip joints in addition to the strength and flexibility of the surrounding musculature.[12] When selecting the individual sub-tests, priority was given to movements previously proposed within the scientific literature as potentially associated with injury likelihood.

The efficacy of screening tests that seek to identify or predict which players will get injured has recently been questioned.[13] In the context of sports-related injuries the idea that a single attribute such as movement quality for example, could be predictive is unlikely.[14] As a result, the ultimate objective of the SIMS will be to investigate whether a *causative relationship* exists between movement quality and injury. Any potential relationship between movement quality and injury is unlikely to be substantial enough to justify the SIMS being considered 'predictive' but it may help inform the content of injury prevention programs by highlighting risk factors.[15]

There is reason to expect that a causative relationship between movement quality and injury may exist since some authors have reported poor FMS™ scores preceding subsequent injury.[8, 16] However, numerous studies utilizing the same movement screening tool have not observed any link.[17] The SIMS may eventually demonstrate a stronger association to injury risk than the FMS™ due to its more explicit scoring criteria (Appendix 2) focusing on specific aspects of each movement. Furthermore, the FMS™ includes movements targeting the upper limbs, which have limited relevance for soccer players, whereas the SIMS concentrates on the lower limbs only.

Before any prospective cohort studies can be conducted using the SIMS its reliability must first be established. The reliability of an assessment tool is of critical importance since it is a pre-requisite for test validity.[18] Therefore, the purpose of the present study was to test the intra- and inter-rater reliability of the SIMS and determine its suitability for use in further research.

## METHODS

### Participants

Twenty-five (11 males, 14 females) healthy, recreationally active university students (age 25.5 ± 4.0 years, height 171 ± 9 cm, weight 64.7 ± 12.6 kg) agreed

to participate in the present study. Inclusion criteria required participants to be aged between 18-40 years of age, free of injury (any physical condition that precluded them from completing the assessment) and recreationally active. Information pertaining to the study protocol and requirements were provided for each participant before written informed consent was collected. The study was approved by the local ethics committee (ref number: 270/15, Ärztekammer des Saarlandes, Saarbrücken, Germany) and conformed to the Declaration of Helsinki.

### Raters

Three raters carried out the SIMS in the present study; all possessed postgraduate sport science qualifications and had previous professional experience delivering movement assessments. In addition, Rater 1 was an accredited strength and conditioning coach with both the United Kingdom Strength and Conditioning Association (UKSCA) and the National Strength and Conditioning Association (NSCA). Prior to the present study all raters conducted pilot testing using the SIMS with 10 participants. The pilot testing incorporated two 2-hour sessions where raters reviewed the test instructions (Appendix 1), the scoring criteria (Appendix 2) and familiarized themselves with the camera positioning (Figure 1). In addition, three more two-hour sessions were conducted where raters practiced scoring video footage and discussed the interpretation of the scoring criteria. In total, rater training amounted to ~12 hours (10 classroom-based and two field-based).

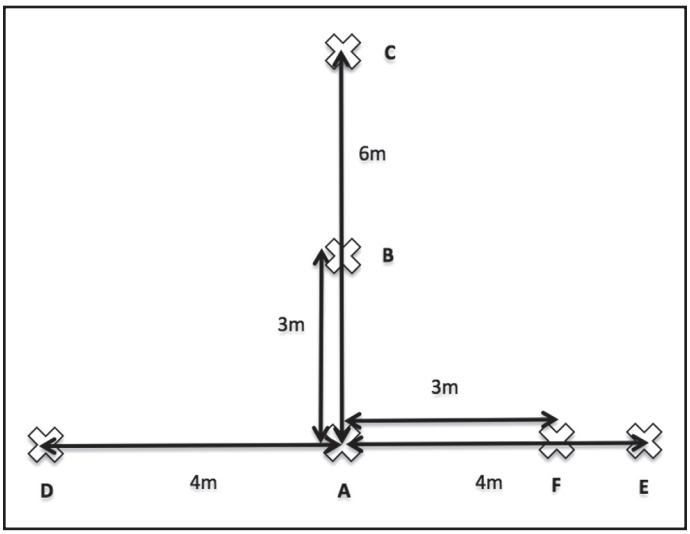



**Figure 1.** *Schematic showing the equipment setup for the SIMS*

For all movements the participants start at A. Anterior reach: measuring tape is fixed to the floor between A and B; Single-leg deadlift: camera at B (portrait) and E (landscape) when standing on right leg, camera at B (portrait) and D (landscape) when standing on left leg; In-line lunge: camera at B (portrait) and E (landscape) when right leg forward, camera at B (portrait) and D (landscape) when left leg forward; Single-leg hop for distance: measuring tape is fixed to the floor between A and C; Tuck jump: taped cross on floor at A (60x60cm), camera at B (portrait) and F (portrait).

### Design

The present study utilized a test-retest design. Participants performed the SIMS on three occasions separated by an average of 3.5 days (minimum one day, maximum seven days). The SIMS contains five sub-tests: the anterior reach (AR), single-leg deadlift (SLDL), in-line lunge (ILL), single-leg hop for distance (SLHD) and tuck jump (TJ) (Figure 2). Raters 1 and 2 scored all participants whereas Rater 3 only scored 15 of the 25 (for reasons unrelated to the study). Raters scored two of the five movements (AR and SLHD) included in the SIMS in real-time on each occasion. The remaining three movements (SLDL, ILL and TJ) were filmed from both the frontal and sagittal planes using iPhone 4S devices (Apple Inc., California, USA) and scored retrospectively. These sub-tests were scored from video footage, as opposed to in real-time; to allow raters to view the movements in slow motion and increase the likelihood of identifying errors. A minimum of one week separated the scoring of participants' filmed movements for occasions one, two and three respectively in an attempt to reduce the risk of rater bias (i.e. remembering the previous scores given). Scores for occasions one, two and three were compared within each rater to investigate 'real-world' intra-rater (inter-occasion) reliability. Scores were also compared between raters for each occasion to assess inter-rater reliability. Rater 1 re-scored the filmed movements for all participants on all occasions six months later to establish the 'pure' intra-rater (intra-occasion) reliability for those movements.

### Soccer Injury Movement Screen (SIMS)

Detailed descriptions of each movement contained within the SIMS and associated scoring criteria are outlined in Appendices 1 and 2. The ILL is the same in its setup as when performed as part of the FMS™ albeit it is scored differently, while the tuck jump is

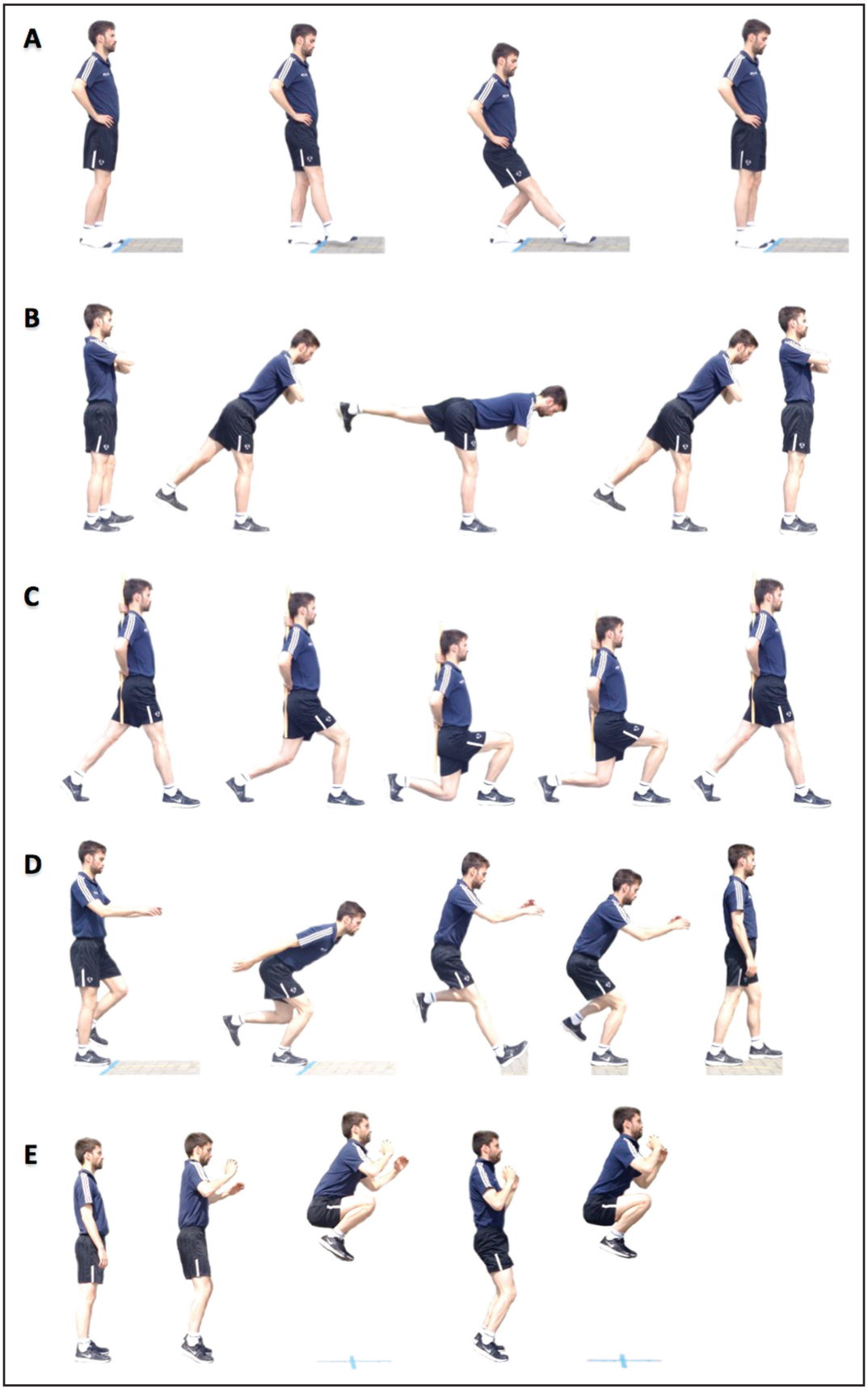



**Figure 2.** *Demonstration cards that were shown to participants along with verbal instructions prior to test execution*
A: anterior reach; B: single-leg deadlift; C: in-line lunge; D: single-leg hop for distance; E: tuck jump.

performed and scored exactly as described by Myer et al.[5, 19] A standardized five minute warm up was completed before each occasion and included dynamic bodyweight exercises (e.g. squats, walking lunges, hamstring walkouts, diagonal hop and holds). The assessments were performed outdoors on a hard, rubberized sports court during summertime in dry temperate weather conditions. Participants were instructed to wear tight fitting sports clothing and the same training shoes on each occasion. The five component movements were performed in sequential order starting with the AR followed by the SLDL, ILL, SLHD and TJ. Prior to each sub-test participants were read the test instructions (Appendix 1) verbatim and shown demonstration cards (Figure 2). Participants were then allowed three practice attempts for each sub-test where any obvious miscommunication or misunderstandings relating to how to execute the movements were clarified. Time to complete the assessment was 10-15 minutes per participant.

Each component movement was scored out of 10 points resulting in a theoretical maximum composite score of 50 when the score from each sub-test is summed. A higher score indicated poorer performance; hence, zero was the theoretical 'best' score while 50 was the 'worst'. The AR and SLHD scoring criteria were objective in nature and were based on reach and jump distance respectively. In contrast, the SLDL, ILL and TJ relied on subjective assessment of movement quality. Raters were allowed to watch the clips of the filmed movements, both in real-time speed and slow motion, as many times as they deemed necessary to make an accurate judgment when scoring.

### Statistical analyses

Descriptive data are presented as means ± standard deviation. Reliability statistics are accompanied with 95% confidence intervals (CI). Data were analysed using R statistics program (R Core Development Team 2014) and MedCalc for Windows, version 16.4.3 (MedCalc Software, Ostend, Belgium). Comparison of composite and individual sub-test scores between male and female participants was performed using the Mann-Whitney U statistic. Cohen's *d* effect size (ES) was also calculated to compare male and female participants and was interpreted as follows: ≤0.2, trivial; 0.21-0.60, small; 0.61-1.2, moderate; 1.21-2.0, large; 2.1-4.0, very large.[20, 21] Two way mixed model intraclass correlation coefficients ($ICC_{3,1}$), weighted kappas (quadratic) and minimal detectable change (MDC) were used to determine the intra- and inter-rater reliability of the composite score. MDC values were calculated at both a 95% and 80% level of confidence in order to provide applied practitioners with the means to identify 'true' changes in test performance. Typically, MDC values are calculated to reflect a 95% confidence interval; however, this results in very conservative estimates of how much a test score has to change to be considered real and may be of limited usefulness in the applied setting where small improvements/decrements in test performance can be meaningful.[22] MDC values at lower levels of confidence (e.g. 80%) can be calculated and are useful to applied practitioners who may be willing to rely on more liberal estimates of test score changes. In addition, weighted kappas (quadratic) were used to determine intra- and inter-rater reliability of each individual subtest. ICC values were interpreted according to the following criteria: <0.40, poor; 0.40-0.59, fair; 0.60-0.74, good; ≥0.75, excellent.[23] Similarly, weighted kappa values were interpreted according to the guidelines outlined by Landis and Koch[24]: <0.00, poor; 0.00-0.20, slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; 0.81-1.00, almost perfect. Alpha was set at $p \leq 0.05$.

## RESULTS

Composite scores were not significantly different between males (18.3) and females (15.3) (Table 1). Only the SLDL scores differed between genders (males = 4.3, females = 1.8) (Table 1).

$ICC_{3,1}$, weighted kappa and MDC values for intra-rater (inter-occasion) reliability are presented in Table 2. Weighted kappa values for the individual subtests ranged from fair to substantial (0.35-0.77). With regard to the composite score, weighted kappa values were interpreted as substantial (0.63-0.68) while the ICCs were classified as good (0.66-0.72) for each rater.

$ICC_{3,1}$ and weighted kappa values for inter-rater reliability are presented in Table 3. Weighted kappa values for the individual subtests ranged from mod-

**Table 1.** *Mean values (reported in arbitrary units) and comparison of test scores between males and females.*

| | Overall (n=25) | Males (n=11) | Females (n=14) | *p-value* | Male vs female effect size (qualitative inference) |
|---|---|---|---|---|---|
| Composite score (mean ± SD) | 16.6 ± 4.9 | 18.3 ± 3.0 | 15.3 ± 5.8 | 0.080 | 0.6 (Small) |
| AR (mean ± SD) | 1.7 ± 1.8 | 2.1 ± 2.3 | 1.4 ± 1.3 | 0.648 | 0.4 (Small) |
| SLDL (mean ± SD) | 2.9 ± 2.1 | 4.3 ± 2.0 | 1.8 ± 1.5 | <0.01 | 1.4 (Large) |
| ILL (mean ± SD) | 2.6 ± 1.5 | 2.5 ± 1.5 | 2.6 ± 1.6 | 0.825 | 0.1 (Trivial) |
| SLHD (mean ± SD) | 4.1 ± 2.3 | 4.2 ± 1.9 | 4.0 ± 2.7 | 0.718 | 0.1 (Trivial) |
| TJ (mean ± SD) | 5.4 ± 1.3 | 5.2 ± 1.0 | 5.5 ± 1.6 | 0.534 | 0.2 (Trivial) |

Test scores drawn from Rater 1 on the third testing occasion. AR= anterior reach, ILL= in-line lunge, SLDL= single-leg deadlift, SLHD= single-leg hop for distance, TJ= tuck jump

**Table 2.** *Summary of intra-rater (inter-occasion) reliability values. Values in brackets represent the 95% confidence intervals.*

| | Weighted kappa | | | | | | $ICC_{3,1}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | AR | SLDL | ILL | SLHD | TJ | Composite score | Composite score | MDC @ 95% confidence | MDC @ 80% confidence |
| Rater 1 | 0.47 (0.17-0.77) | 0.77 (0.67-0.87) | 0.64 (0.52-0.77) | 0.44 (0.26-0.61) | 0.58 (0.43-0.73) | 0.68 (0.54-0.81) | 0.71 (0.52-0.85) | 7.0 | 4.5 |
| Rater 2 | 0.46 (0.22-0.69) | 0.68 (0.55-0.81) | 0.48 (0.30-0.66) | 0.35 (0.15-0.55) | 0.58 (0.44-0.72) | 0.64 (0.49-0.80) | 0.72 (0.54-0.85) | 7.5 | 4.9 |
| Rater 3 | 0.39 (0.02-0.77) | 0.68 (0.55-0.81) | 0.63 (0.49-0.77) | 0.36 (0.11-0.61) | 0.45 (0.26-0.65) | 0.63 (0.45-0.80) | 0.66 (0.38-0.86) | 6.7 | 4.4 |

AR= anterior reach, ICC= intra-class correlation coefficient, ILL= in-line lunge, MDC= minimum detectable change, SLDL= single-leg deadlift, SLHD= single-leg hop for distance, TJ= tuck jump

**Table 3.** *Summary of inter-rater reliability values (between all three raters). Values in brackets represent the 95% confidence intervals.*

| | Weighted kappa | | | | | | $ICC_{3,1}$ |
|---|---|---|---|---|---|---|---|
| | AR | SLDL | ILL | SLHD | TJ | Composite score | Composite score |
| Occasion 1 | 0.83 (0.72-0.95) | 0.51 (0.35-0.66) | 0.71 (0.58-0.85) | 0.84 (0.69-1.00) | 0.60 (0.40-0.81) | 0.78 (0.68-0.88) | 0.79 (0.58-0.92) |
| Occasion 2 | 0.76 (0.62-0.90) | 0.48 (0.29-0.66) | 0.70 (0.56-0.84) | 0.91 (0.85-0.97) | 0.43 (0.18-0.68) | 0.81 (0.71-0.90) | 0.86 (0.70-0.95) |
| Occasion 3 | 0.59 (0.33-0.84) | 0.64 (0.50-0.79) | 0.58 (0.41-0.75) | 0.91 (0.86-0.97) | 0.50 (0.35-0.65) | 0.79 (0.70-0.87) | 0.79 (0.58-0.92) |

AR= anterior reach, ICC= intra-class correlation coefficient, ILL= in-line lunge, SLDL= single-leg deadlift, SLHD= single-leg hop for distance, TJ =tuck jump

erate to almost perfect (0.43-0.91). With regard to the composite score weighted kappa values ranged from substantial to almost perfect (0.78-0.81) while the ICCs were classified as excellent (0.79-0.86) for all three occasions.

Weighted kappa scores for 'pure' intra-rater (intra-occasion) reliability are presented in Table 4. The kappa values were evaluated as almost perfect for the SLDL (0.90) and ILL (0.85) while the TJ value was interpreted as substantial (0.73).

## DISCUSSION

Overall, the present results indicate sufficient reliability for the SIMS to be considered useful for further research and applied practitioners alike. The intra-rater reliability of the SIMS composite score was classed as substantial and good for all raters based upon the weighted kappa and ICC scores respectively (Table 2). The MDC values calculated

**Table 4.** *Summary of intra-rater (intra-occasion) reliability values for video-taped movements. Values in brackets represent the 95% confidence intervals.*

| | Weighted kappa | | |
|---|---|---|---|
| | SLDL | ILL | TJ |
| Rater 1 | 0.90 (0.86-0.95) | 0.85 (0.80-0.91) | 0.73 (0.62-0.83) |

ILL= in-line lunge, SLDL= single-leg deadlift, TJ= tuck jump

at an 80% level of confidence demonstrate that if a one-point increase or decrease in each sub-test were observed a 'real' change in composite score would have likely occurred. The inter-rater reliability was classified as substantial to almost perfect when considering the weighted kappa values and excellent according to the ICCs (Table 3). The SLDL sub-test was the only movement where a discrepancy in scores between males and females was apparent (Table 1). Male participants regularly cited hamstring inflexibility as a limiting factor during this task whereas female participants rarely mentioned this. Females generally display superior hamstring flexibility as compared to men.[25] This difference in hamstring flexibility between males and females may potentially explain the gender difference in SLDL score observed in the present study.

The AR portion of the Y-balance test has previously been investigated as a risk factor with limb asymmetry >4 cm equating to a 2.3 – 2.7 times greater likelihood of non-contact injury among basketball and track and field athletes.[26, 27] The scoring criteria used in this assessment (Appendix 2) required the rater to assign a score (0 – 10) based on the difference in reach distance between limbs. The reason for limiting the scoring range to a maximum of 10 points (a reach asymmetry of ≥10 cm) was to maintain equal weighting between all five sub-tests (each of which was scored out of 10). The scoring criteria were clearly objective for this sub-test and therefore did not directly assess movement quality. However, it was decided that the AR warranted inclusion in the SIMS regardless of not directly assessing movement quality, due to the promising evidence surrounding its relationship to injury.[26, 27] The test reflects a number of physical qualities including neuromuscular control, strength and ankle stability: all of which are likely contributors to movement quality.[1, 26, 27] Therefore, while this sub-test did not assess movement quality directly the variable that was measured (difference in reach distance) is likely a reasonable surrogate marker. Ankle injuries occur frequently within soccer therefore the anterior reach may be a promising tool for highlighting increased risk of such events.[28] The intra-rater weighted kappa values for the AR ranged from fair to moderate (Table 2). In contrast, the inter-rater values ranged from moderate to almost perfect (Table 3). The difference between the intra- and inter-rater weighted kappa values suggests that the scoring criteria were clear but that a large proportion of the variation in the test scores stemmed from the participants and/or the influence of time between testing occasions. As such, additional participant familiarization with the test may help improve the intra-rater reliability.

While the SLDL is multifaceted in its demands, eccentric strength and flexibility of the hamstrings are clearly primary aspects of the movement due to the flexion of the hip with an extended knee on the standing leg. Both eccentric strength and flexibility of the hamstrings have been proffered as injury risk factors within soccer players.[29, 30] Hence, the ability to perform the SLDL with a high degree of movement quality may indicate proficiency in these important attributes (hamstring flexibility and eccentric strength). The intra-rater SLDL weighted kappa values for each rater represented substantial agreement (Table 2) while the inter-rater reliability values ranged from moderate to substantial (Table 3). These findings suggest that while raters were very consistent in their scoring of the SLDL within themselves there is opportunity for improvement in the between-rater agreement. Such a scenario is somewhat inevitable when considering subjective scoring criteria; however, more detailed guidelines on what constitutes a movement 'error' may help improve consensus between raters in the future.

The ILL, or split squat, is a widely used exercise within soccer both during warm-up routines and resistance training sessions.[31, 32] According to Cook et al.[34] the ILL focuses on the "stresses simulated during rotational, decelerating and lateral type movements". All of these movement patterns are frequently observed during soccer match play.[34] The ability to perform this exercise correctly is important to ensure players do not use compensatory movements that potentially cause or exacerbate acute and overuse injuries. When performing the ILL the same test setup was used as with the FMS™; however, the scoring criteria utilized in the current research (Appendix 2) differed.[33] The alternative scoring criteria were employed with the intention of explicitly outlining the potential movement flaws and hence enhancing clinical usefulness of the results. Both intra- and

inter-rater reliability of the ILL ranged from moderate to substantial (Tables 2 & 3). The weighted kappa values reported in the present investigation are in keeping with those observed in studies of the FMS™ version of the ILL.[35-37] The more detailed scoring criteria adopted by the SIMS as compared with the FMS™ did not appear to adversely affect the reliability yet will provide practitioners with a clearer indication of where any potential movement dysfunction originates from.

It is important for soccer-specific movement assessments to incorporate explosive actions such as jumping and landing since they occur frequently during match play and often precede serious injury.[34, 38] While bilateral, vertical drop jumps have long been used for injury risk stratification[39, 40] many explosive soccer-specific actions are unilateral in nature and involve horizontal as well as vertical displacement (for example: kicking, changing direction and landing after a header).[34] The scoring criteria for the SLHD were objective and incorporated both the jump distance and the between limb difference in jump distance (Appendix 2) with each of these aspects weighted equally. The precise distances that characterized the different scoring ranges were based on pilot testing conducted with recreationally active university students and therefore may not be applicable to professional or youth soccer players. Revised criteria may need to be established for higher-level athletes. The authors opted for objective, as opposed to subjective, scoring in this instance due to recent evidence suggesting jump distance as a risk factor for non-contact hamstring injury.[41] While the intra-rater weighted kappa values ranged from fair to moderate the inter-rater values indicated almost perfect agreement between raters (Tables 2 & 3). The discrepancy between the intra- and inter-rater weighted kappa values suggests that a large proportion of the variation in the test scores stemmed from the participants and/or the influence of time between testing occasions rather than the application of the scoring criteria per se.

Allowing more jump attempts may increase the likelihood of maximum jump distance being reached and a plateau in performance occurring, which may in turn help improve reliability. On 32 of the 75 SLHD tests scored by Rater 1, (25 participants on three occasions) participants recorded their best jump distance (for that occasion) on their last attempt. Similarly, 15 of the 25 participants recorded their best jump distances overall on testing occasion 3. In addition, 12 of the 25 participants scored by Rater 1 recorded their best between limb difference score on their third testing occasion. This demonstrates that incorporating a number of familiarization sessions on multiple days prior to testing may improve reliability for the same reasons highlighted previously (plateauing of performance). However, it should be remembered that the more attempts allowed and the more familiarization sessions performed the greater the potential for fatigue to influence test performance and the less practically feasible the assessment may become. There may be a trade-off between improved reliability and the feasibility of using the SIMS as a screening tool in the applied environment. A recent systematic review by Hegedus et al.[42] assessed the methodological quality of studies exploring the reliability and validity of commonly used field-expedient screening tests such as the SLHD. They found no studies of satisfactory methodological quality reporting the reliability of the SLHD precluding comparison of the current results to previous findings.

The TJ assessment has been proposed as a field-expedient assessment of lower limb neuromuscular control.[19] It is unique as an assessment of movement quality since it requires the participant to continuously perform plyometric vertical jumps for 10 seconds.[19] While it is unlikely a player would replicate this precise activity during match-play the taxing nature of the test means it is likely to expose potentially injurious lower-limb movement patterns (particularly those associated with the onset of fatigue) that other, typically lower intensity assessments may not highlight. It has been suggested as a particularly useful tool for highlighting knee valgus movement during landing, which has been proposed as a risk factor for anterior cruciate ligament (ACL) injury.[19, 43] Considering the long-term sequelae associated with ACL injury the authors judged the TJ worthy of inclusion in the SIMS.[44, 45] Both the intra- and inter-rater weighted kappa values represented moderate agreement within and between raters (Tables 2 & 3). While this indicates accept-

able reliability the weighted kappa values calculated are lower than previously reported by Myer et al.[19] However, Myer et al.[19] only assessed 10 participants and so raters may have remembered the previous scores given, leading to recall bias. In addition, they scored the same video footage twice as opposed to scoring participants on two separate occasions. The scoring criteria (Appendix 2) are inherently subjective but reliability may be improved by adding some objective guidelines to certain scoring items. For example, one of the scoring items asks: "was there a pause between jumps"? This could potentially be changed to: "was there a pause, lasting longer than one second (or another defined time period), between jumps"? Such amendments may improve consistency of scoring within and between raters. However, future research is needed to assess the difference in reliability when objective instructions are given compared with when they are not.

In an effort to separate some of the sources of variation within the test-retest design, one rater scored all the filmed movements (SLDL, ILL and TJ) from each testing occasion twice. This removed the influence of variation in test performance stemming from the participants and revealed the 'pure' intra-rater, or intra-occasion, reliability. The weighted kappa values for the SLDL and ILL represented almost perfect agreement while the score for the TJ indicated substantial reliability (Table 4). These higher weighted kappa values (as compared to those reported in Table 2) are not surprising since they reflect only the variation in scoring associated with the rater. These results suggest that improvements in the 'real world' intra-rater reliability are more likely to arise from aspects related to the participants rather than the raters. Bearing this in mind, future strategies aimed at improving the intra-rater reliability of the SIMS further may include extended participant familiarization with the test and allowing them to read the scoring criteria. Explicitly explaining the scoring criteria for the FMS™ to participants elicited improved scores.[46] This suggests that ambiguity related to what is being asked of participants during movement screening may influence their test execution and potentially contribute to variation in performance.

A number of limitations should be considered when interpreting the results of the present study. Perhaps most importantly, the pilot testing conducted to establish the scoring ranges for the SLHD (Appendix 2) were based on recreationally active university students' scores. As such, it may be necessary to revise this aspect of the scoring criteria in the future if the SIMS is used with professional soccer players. Similarly, if the SIMS were to be utilized with youth soccer players then amendments to the scoring criteria may be necessary. In addition, the results presented here are from only 25 participants, which, is a relatively modest sample size for assessing reliability according to Terwee et al[47]; however, the scores from three trials were included, rather than the usual two in an effort to improve the credibility of the conclusions. Furthermore, the raters represented a homogenous group. All were PhD students with postgraduate degrees in sport science. Further research may be needed to assess the reliability of the SIMS when conducted by other groups of raters, for example, undergraduate students or sports coaches.

## CONCLUSIONS

Until now, no movement screen has been developed specifically for use among soccer players. The SIMS composite score demonstrated good to excellent intra- and inter-rater reliability. However, the intra-rater reliability of the individual sub-tests ranged from fair to substantial indicating scope for further improvement. Establishing the reliability of the SIMS is a prerequisite for further research seeking to investigate the relationship between test score and subsequent injury. The present results indicate at least acceptable reliability for this purpose.

## REFERENCES


1. McCunn R, aus der Fünten K, Fullagar HHK, et al. Reliability and association with injury of movement screens: A critical review. *Sports Med*. 2016;46(6):763-781.
2. McCall A, Carling C, Nedelec M, et al. Risk factors, testing and preventative strategies for non-contact injuries in professional football: Current perceptions and practices of 44 teams from various premier leagues. *Br J Sports Med*. 2014;48(18):1352-1357.
3. Reid DA, Vanweerd RJ, Larmer PJ, et al. The inter and intra rater reliability of the Netball Movement Screening Tool. *J Sci Med Sport*. 2015;18(3):353-357.
4. Padua DA, Marshall SW, Boling MC, et al. The Landing Error Scoring System (LESS) is a valid and

reliable clinical assessment tool of jump-landing biomechanics: The JUMP-ACL study. *Am J Sports Med*. 2009;37(10):1996-2002.

5. Cook G, Burton L, Hoogenboom B. Pre-participation screening: The use of fundamental movements as an assessment of function – part 1. *N Am J Sports Phys Ther*. 2006;1(2):62-72.
6. Moran RW, Schneiders AG, Major KM, et al. How reliable are Functional Movement Screening scores? A systematic review of rater reliability. *Br J Sports Med*. 2016;50(9):527-536.
7. McGill S, Frost D, Lam T, et al. Can fitness and movement quality prevent back injury in elite task force police officers? A 5-year longitudinal study. *Ergonomics*. 2015;58(10):1682-1689.
8. Lisman P, O'Connor FG, Deuster PA, et al. Functional movement screen and aerobic fitness predict injuries in military training. *Med Sci Sports Exerc*. 2013;45(4):636-643.
9. Kiesel K, Plisky PJ, Voight ML. Can serious injury in professional football be predicted by a preseason functional movement screen? *N Am J Sports Phys Ther*. 2007;2(3):147-158.
10. Padua DA, DiStefano LJ, Beutler AI, et al. The Landing Error Scoring System as a screening tool for an anterior cruciate ligament injury-prevention program in elite-youth soccer athletes. *J Athl Train*. 2015;50(6):589-595.
11. Parsonage JR, Williams RS, Rainer P, et al. Assessment of conditioning-specific movement tasks and physical fitness measures in talent identified under 16-year-old rugby union players. *J Strength Cond Res*. 2014;28(6):1497-1506.
12. Ekstrand J, Hägglund, Waldén M. Injury incidence and injury patterns in professional football: the UEFA injury study. *Br J Sports Med*. 2011;45(7):553-558.
13. Bahr R. Why screening tests to predict injury do not work – and probably never will…: A critical review. *Br J Sports Med*. 2016;50(13):776-780.
14. Cook C. Predicting future physical injury in sports: It's a complicated dynamic system. *Br J Sports Med*. 2016 [Epub ahead of print].
15. McCunn R, Meyer T. Screening for risk factors: If you liked it then you should have put a number on it. *Br J Sports Med*. 2016 [Epub ahead of print].
16. Garrison M, Westrick R, Johnson MR, et al. Association between the functional movement screen and injury development in college athletes. *Int J Sports Phys Ther*. 2015;10(1):21-28.
17. Warren M, Smith CA, Chimera NJ. Association of the functional movement screen with injuries in division I athletes. *J Sport Rehabil*. 2015;24(2):163-170.
18. Batterham AM, George KP. Reliability in evidence-based clinical practice: A primer for allied health professionals. *Phys Ther Sport*. 2003;4(3):122-128.
19. Myer GD, Ford KR, Hewett TE. Tuck jump assessment for reducing anterior cruciate ligament injury risk. *Athl Ther Today*. 2008;13(5):39-44.
20. A new view of statistics: Effect magnitudes. http://www.sportsci.org/resource/stats/effectmag.html Published 2002. Updated August 7, 2006. Accessed August 1, 2016.
21. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155-159.
22. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19(1):231-240.
23. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assessment*. 1994;6(4):284-290.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
25. Marshall PW, Siegler JC. Lower hamstring extensibility in men compared to women is explained by differences in stretch tolerance. *BMC Musculoskelet Disord*. 2014;15:223.
26. Smith CA, Chimera NJ, Warren M. Association of y balance test reach asymmetry and injury in division I athletes. *Med Sci Sports Exerc*. 2015;47(1):136-141.
27. Plisky PJ, Rauh MJ, Kaminski TW, et al. Star excursion balance test as a predictor of lower extremity injury in high school basketball players. *J Orthop Sports Phys Ther*. 2006;36(12):911-919.
28. Walls RJ, Ross KA, Fraser EJ, et al. Football injuries of the ankle: A review of injury mechanisms, diagnosis and management. *World J Orthop*. 2016;7(1):8-19.
29. Timmins RG, Bourne MN, Shield AJ, et al. Short biceps femoris fascicles and eccentric knee flexor weakness increase the risk of hamstring injury in elite football (soccer): A prospective cohort study. *Br J Sports Med*. 2015 [Epub ahead of print].
30. van Beijsterveldt AM, van de Port IG, Vereijken AJ, et al. Risk factors for hamstring injuries in male soccer players: A systematic review of prospective studies. *Scand J Med Sci Sports*. 2013;23(3):253-262.
31. Bizzini M, Dvorak J. FIFA 11+: An effective programme to prevent football injuries in various player groups worldwide – a narrative review. *Br J Sports Med*. 2015;49(9):577-579.
32. Owen AL, Wong DP, Dellal A, et al. Effect of an injury prevention program on muscle injuries in elite professional soccer. *J Strength Cond Res*. 2013;27(12):3275-3285.

33. Cook G, Burton L, Hoogenboom BJ, et al. Functional movement screening: The use of fundamental movements as an assessment of function – part 1. *Int J Sports Phys Ther*. 2014;9(3):396-409.
34. Stølen T, Chamari K, Castagna C, et al. Physiology of soccer: An update. *Sports Med*. 2005;35(6):501-536.
35. Teyhen DS, Shaffer SW, Lorenson CL, et al. The Functional Movement Screen: A reliability study. *J Orthop Sports Phys Ther*. 2012;42(6):530-540.
36. Schneiders AG, Davidsson A, Hörman E, et al. Functional movement screen normative values in a young, active population. *Int J Sports Phys Ther*. 2011;6(2):75-82.
37. Minick KI, Kiesel KB, Burton L, et al. Interrater reliability of the functional movement screen. *J Strength Cond Res*. 2010;24(2):479-486.
38. Waldén M, Krosshaug T, Bjørneboe J, et al. Three distinct mechanisms predominate in non-contact anterior cruciate ligament injuries in male professional football players: A systematic video analysis of 39 cases. *Br J Sports Med*. 2015;49(22):1452-1460.
39. Nilstad A, Andersen TE, Kristianslund E, et al. Physiotherapists can identify female football players with high knee valgus angles during vertical drop jumps using real-time observational screening. *J Orthop Sports Phys Ther*. 2014;44(5):358-365.
40. Ekegren CL, Miller WC, Celebrini RG, et al. Reliability and validity of observational risk screening in evaluating dynamic knee valgus. *J Orthop Sports Phys Ther*. 2009;39(9):665-674.
41. Goossens L, Witvrouw E, Vanden Bossche L, et al. Lower eccentric hamstring strength and single leg hop for distance predict hamstring injury in PETE students. *Eur J Sport Sci*. 2015;15(5):436-442.
42. Hegedus EJ, McDonough S, Bleakley C, et al. Clinician-friendly lower extremity physical performance measures in athletes: A systematic review of measurement properties and correlation with injury, part 1. The tests for knee function including the hop tests. *Br J Sports Med*. 2015;49(10):642-648.
43. Ford KR, Myer GD, Hewett TE. Valgus knee motion during landing in high school female and male basketball players. *Med Sci Sports Exerc*. 2003;35(10):1745-1750.
44. Waldén M, Hägglund M, Magnusson H, et al. ACL injuries in men's professional football: A 15-year prospective study on time trends and return-to-play rates reveals only 65% of players still play at the top level 3 years after ACL rupture. *Br J Sports Med*. 2016;50(12):744-750.
45. Øiestad BE, Engebretsen L, Storheim K, et al. Knee osteoarthritis after anterior cruciate ligament injury: A systematic review. *Am J Sports Med*. 2009;37(7):1434-1443.
46. Frost DM, Beach TA, Callaghan JP, et al. FMS scores change with performers' knowledge of the grading criteria – Are general whole-body movement screens capturing "dysfunction"? *J Strength Cond Res*. 2015;29(11):3037-3044.
47. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21(4):651-657.

**Appendix 1.** *Description of the Soccer Injury Movement Screen (SIMS).*

| Movement name | Rationale/perceived usefulness | Instructions |
|---|---|---|
| Pre-assessment | N/A | "For each exercise you have three practice attempts and three scored attempts on each leg. In the case of the tuck jump you have three practice jumps followed by the scored 10 second effort." |
| Anterior reach | - Provides an indication of ankle mobility (dorsiflexion)<br>- Highlights limb asymmetry (ankle mobility and/or leg strength)<br>- Provides an indication of single-leg control (e.g. motor control and balance) | "Remove your shoes. Place the big toe of your standing leg so it is touching the back of the taped line. Place hands on your hips. Reach the toes of the other leg as far along the measuring tape as possible – hovering around 5 centimeters off the ground. You must keep your standing foot in contact with the floor throughout, e.g. you cannot rise up on to your toes. Try to hover at the point of maximal reach for a couple of seconds to allow scoring. You must return to the start position for the attempt to be counted. Likewise, you must maintain balance throughout each attempt for the score to be recorded."* |
| Single-leg deadlift | - Provides an indication of ability to simultaneously flex and extend at the hip with extended knees while maintaining neutral spinal alignment<br>- Provides an indication of hamstring flexibility<br>- Provides an indication of single-leg control (e.g. motor control and balance) | "Put your shoes back on. Tuck your t-shirt into your shorts. Stand on the middle of the cross, taped on the floor, and cross arms over your chest. Imagine a straight line between your head and your right heel. Try to hinge at the hip while keeping that line straight until parallel to the floor. Try to keep your standing leg (left) extended. Return to the start position with both feet touching the floor between each repetition." Switch the words 'right' and 'left' when instructing the participant when testing the other side. |

* If available, a slider device (e.g. Y Balance Test Kit™) can be used to perform the anterior reach.

| Movement name | Rationale/perceived usefulness | Instructions |
|---|---|---|
| In-line lunge | - Provides an indication of ability to simultaneously flex and extend at the hip with flexed knees while maintaining neutral spinal alignment<br>- Provides an indication of lower limb motor control and balance | As per instructions from Functional Movement Screen (Cook et al. 2006a) (see reference list for full article details). "Place your left toes so they are touching the back of the taped line. Place the heel of your right foot xx centimeters (as marked by instructor)** directly in front of your left foot. Hold the dowel behind your back gripping it with your left hand at your neck and your right hand at your lower back. Make sure the dowel is touching your head, upper back and buttocks. While maintaining an upright posture, descend into a lunge touching your left knee to the floor. Maintain contact with the dowel at the head, upper back and bum throughout. Return to the start position with knees fully extended between each repetition." Switch the words 'right' and 'left' when instructing the participant when testing the other side. |
| Single-leg hop for distance | - Provides an indication of lower-limb unilateral power<br>- Highlights limb asymmetry (lower-limb power and/or ankle stability and/or lower-limb eccentric strength)<br>- Provides an indication of single-leg control | "Place the toes of the jumping leg so they are touching the back of the taped line. Jump as far as you can while still able to stick the landing on the same leg and hold your position to allow measurement. You must record three successful scored jumps on each leg and you will receive as many attempts as necessary to achieve this." |
| Tuck jump | - Allows quick assessment of bilateral knee control during plyometric activity<br>- Highlights limb asymmetry (lower-limb power and/or hip mobility) | As per instructions from Myer et al. (2008) (see reference list for full article details). "Stand on the middle of the cross taped on the floor with feet shoulder width apart. Upon signal from the tester, perform continuous vertical jumps on the spot for 10 seconds making sure to lift your knees towards your chest so that your upper thighs are parallel with the floor each time. Try to perform as many jumps as possible." |

** Foot placement is determined by measuring the distance from the floor to the tibial tuberosity (shin length).

## APPENDIX 2. SCORING CRITERIA

### General rater instructions

Record each participant's height, weight and tibial tuberosity height (distance from the floor to their tibial tuberosity). If a participant cannot physically perform any test due to pain then they should be considered injured, this should be reported to the relevant club staff members and the test should be postponed.

### Scoring guidelines for the anterior reach and single-leg hop for distance (objective assessments)

### Anterior reach

Measure the distance (in centimeters) from the start line to the most distal part of the foot of the reaching leg. Round to the nearest centimeter. Three repetitions are performed on each leg and reach distance should be recorded for each attempt. The maximum reach distances achieved by each leg should be used to calculate the difference between left and right. The maximum theoretical score achievable is 10 and this would represent a 'poor' score. In contrast, the theoretical minimum score is zero and this would represent a 'good' score.

| Difference in reach distance (cm) between legs | Test score |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| 8 | 8 |
| 9 | 9 |
| ≥10 | 10 |

### Single-leg hop for distance

Measure the distance (in centimeters) from the start line to the heel of the jumping/landing leg. Round to the nearest centimeter. Three repetitions are performed on each leg and jump distance should be recorded for each attempt. Both jump distance and limb symmetry are taken into account when assigning a test score. The maximum jump distance achieved on each leg should be summed and used to calculate the score. Combine the scores for jump distance and jump symmetry to produce the final score out of 10. The maximum theoretical score achievable is 10 and this would represent a 'poor' score. In contrast, the theoretical minimum score is zero and this would represent a 'good' score.

| Sum of right and left best jump distances (cm) | | Test score |
|---|---|---|
| Males: | Females: | |
| < 320 | < 220 | 5 |
| 321-340 | 221-240 | 4 |
| 341-360 | 241-260 | 3 |
| 361-380 | 261-280 | 2 |
| 381-400 | 281-300 | 1 |
| > 400 | > 300 | 0 |

| Difference between best right and left jumps (cm) | Test score |
|---|---|
| > 20 | 5 |
| 17-20 | 4 |
| 13-16 | 3 |
| 9-12 | 2 |
| 4-8 | 1 |
| < 4 | 0 |

### Scoring guidelines for the single-leg deadlift, in-line lunge and tuck jump (subjective assessments)

- If an error occurs once and the rater judges it to be egregious then it should be scored as an error.
- If an error (but only to a minor extent) is observed once then it should not be scored.
- If the same error (but only to a minor extent) is observed twice then it should be scored as an error.

Defining specifically what constitutes "minor extent" or "egregious" is not possible. These judgments are left to the discretion of each individual rater. An important consideration is that raters are consistent in their judgments within themselves.

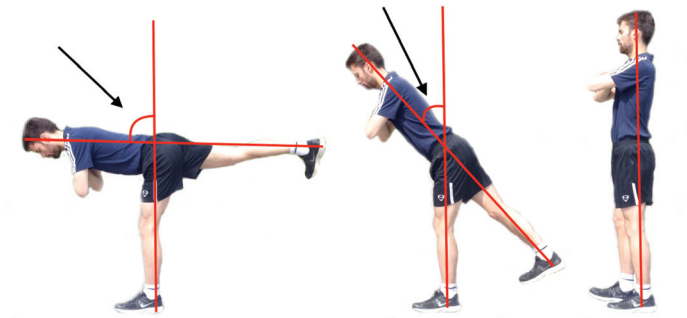
### Single-leg deadlift

The score for this test is based on the 'movement quality' criteria outlined below. Three repetitions are performed on each leg. The maximum theoretical score achievable is 10 and this would indicate 'poor' movement quality. In contrast, the theoreti-

cal minimum score is zero and this would indicate 'good' movement quality. Both legs are scored and the average of both right and left scores is assigned to the individual.

| Item # | | |
|---|---|---|
| 1 | Is external hip rotation (standing leg) visible? | Yes = 1 No = 0 |
| 2 | Does lumbar spine remain neutral? | Yes = 0 No = 1 |
| 3 | Does thoracic spine remain neutral? | Yes = 0 No = 1 |
| 4 | Does knee of raised leg remain extended throughout? | Yes = 0 No = 1 |
| 5 | Is upper and lower body movement synchronized? | Yes = 0 No = 1 |
| 6 | Is footprint maintained? | Yes = 0 No = 1 |
| 7 | Is hip abduction (standing leg) present? | Yes = 1 No = 0 |
| 8 | Does the standing leg knee remain extended throughout? | Yes = 0 No = 1 |
| 9 | Parallel to floor position achieved? Parallel (90˚) = 0, 89˚-45˚ = 1, < 45˚ = 2 (all relative to the stance leg hip flexion angle) | |

In relation to item #9 – the angle being assessed is displayed in the following diagram:



### In-line lunge

The score for this test is based on the 'movement quality' criteria outlined below. Three repetitions are performed on each side. The maximum theoretical score achievable is eight and this would indicate 'poor' movement quality. In contrast, the theoretical minimum score is zero and this would indicate 'good' movement quality. Both legs are scored and the average of both right and left scores is assigned to the individual. To generate a score out of 10 multiply the fractional score out of eight by 10 e.g. if an individual displays four out of eight possible errors then the score out of 10 is: (4/8)x10 = 5. The reason for generating a score out of 10 is to maintain the same weighting between the five sub-tests.

| Item # | | |
|---|---|---|
| 1 | Does dowel remain vertical in frontal plane throughout? | Yes = 0 No = 1 |
| 2 | Does torso rotation (transverse plane) occur? | Yes = 1 No = 0 |
| 3 | Does dowel remain vertical in sagittal plane throughout? | Yes = 0 No = 1 |
| 4 | Does back knee touch the floor? | Yes = 0 No = 1 |
| 5 | Does heel of front foot lift off the floor? | Yes = 1 No = 0 |
| 6 | Is footprint maintained throughout? | Yes = 0 No = 1 |
| 7 | Are the three dowel contact points with body maintained? | Yes = 0 No = 1 |
| 8 | Does knee valgus occur during the movement? | Yes = 1 No = 0 |

### Tuck jump

Mark a cross on the floor using tape (two 60cm strips that intersect). The score for this test is based on the 'movement quality' criteria outlined below. The maximum theoretical score achievable is 10 and this would indicate 'poor' movement quality. In contrast, the theoretical minimum score is zero and this would indicate 'good' movement quality. Myer et al. (2008) created the tuck jump assessment and any further clarification on scoring procedures can be sought from their original article (see reference list for full article details).

| Item # | | |
|---|---|---|
| 1 | Was there knee valgus at landing? | Yes = 1 No = 0 |
| 2 | Do thighs reach parallel (peak of jump)? | Yes = 0 No = 1 |
| 3 | Were thighs equal side-to-side (during flight)? | Yes = 0 No = 1 |
| 4 | Was foot placement shoulder width apart? | Yes = 0 No = 1 |
| 5 | Was foot placement parallel (front to back)? | Yes = 0 No = 1 |
| 6 | Was foot contact timing equal? | Yes = 0 No = 1 |
| 7 | Was there excessive contact landing noise? | Yes = 1 No = 0 |
| 8 | Was there a pause between jumps? | Yes = 1 No = 0 |
| 9 | Did technique decline prior to 10 seconds? | Yes = 1 No = 0 |
| 10 | Were landings in same footprint (within taped cross)? | Yes = 0 No = 1 |